Zpravodaj Československého sdružení uživatelů TeXu

Janka Chlebíková

Ako rozděliť (slovo) Československo

Zpravodaj Československého sdružení uživatelů TeXu, Vol. 1 (1991), No. 4, 10–13

Persistent URL: http://dml.cz/dmlcz/148815

Ako rozdeliť (slovo) Československo.

Každý, kto používal TEX na písanie v slovenčine či češtine v dobe, kedy TEX ešte nebol prispôsobený na naše československé (česko-slovenské, české a slovenské)[2] pomery, musel sa vysporiadať s dvoma nákladnými problémami:

1. Akcentované znaky – špeciálne ľ, ď, ť, (`\v t` totiž láva ť, s čím sa nemožno uspokojiť). Nemožno tiež zabudnúť na „naše“ úvodzovky, ktoré sa líšia od “anglických”.
2. Rozdeľovanie slov – neprispôsobené pôvodné verzie TEXu rozdeľujú slová podľa pravidiel anglickej gramatiky, vôbec nerozdeľujú akcentované slová a ani ich nedovoľujú pridávať do slovníka výnimiek `\hyphenation{...}`.

Prvý problém je možno v súčasnosti veľmi pohodlne riešiť (odhliadnuc od pamäťových nárokov) napr. použitím 8-bitových fontov od P. Nováka. Pred dokončením sú už i 8-bitové DC-fonty, obsahujúce vätšinu znakov z najrozšírenejších európskych jazykov (i naše ť, ď, ľ, f). Tieto fonty by sa mali stať medzinárodným štandardom. Navyše všeobecne rozšírený emTEX poskytuje pohodlnú možnosť konverzie akcentovaných znakov do ľubovoTných kódov, takže editovanie si môžeme spríjemniť napr. použitím známeho programu od Kamenických. Podľa mojich osobných skúseností ani posledná verzia PC TEXu neumožňuje tak jednoduchú konverziu ako emTEX (Američanov až nataľko netrápia naše akcentované znaky).

S druhým problémom sa musí každý jazyk vysporiadať sám vytvorením vlastného „vzorkovníka“. Pre češtinu ho vytvoril Láďa Lhotka a české delenie funguje celkom úspešne. I napriek tomu, že naše jazyky sú blízke, gramatické jemnôstky ma viedli k pokusom s vytváraním slovenských vzorov pre delenie (a nie poslovenčením českých, čo by sa na prvý pohľad mohlo zdať jednoduchšie).

Rozdeľovanie slov na počítači sa nedá urobiť dokonale. Správnosť rozdelenia často závisí od kontextu, napr. na-drobiť a nad-robiť. Autorom rozdeľovacieho algoritmu pre TEX, ako i „vzorkovníka“ pre angličtinu je Frank M. Liang. Svoje výsledky (hlavne problém vytváranie vzorov) zhrnul do Ph. D. Thesis (1983) na Standfordovej Univerzite. Navrhnutá metóda je dostatočne rýchla a láva možnosť pomocou iných vzorov učiť TEX rozdeľovať i v iných jazykoch (napr. v slovenčine alebo v češtine). Kvalita delenia závisí potom už len od nás, kolko času venujeme príprave vzorov.

TEX sa pri delení slova najskôr pozrie do slovníka výnimiek `\hyphenation{...}`. Ak tam slovo nájde, rozdelí ho spôsobom, ako mu predpisujeme, napr. `\hyphenation{výnim-ka}`. V takomto prípade už nepoužíva delenie pomocou „vzorkovníka“ (takže uvedením slova do bez rozdeľovníkov možno zabrániť rozdeľovaniu slova. v celom dokumente). Obdobne je tomu i u slov, v ktorých sa zadáva delenie explicitne pomocou `\-`, napr. po`\-`mlč`\-`ka.

Pokiaľ nenastane ani jedna zo spomínaných možností, spustí Liangov algoritmus. Ukážme si, ako bude rozdelené slovo `Czechoslovakia` s anglickými vzormi zo súboru *enhyph.tex* (štandardne distribuovaného s CSTEXom). Pri rozdeľovaní slov velké a malé písmen nehrajú rolu, preto TEX najskôr „preloží“ slovo do malých písmen (samozrejme tento preklad musí byť presne definovaný). Na začiatok a koniec slova pridá extra znaky – bodky, ktoré budú v ďalšom signalizovať, že sa jedná o špeciálnu časť slova (začiatok a koniec).

[2] Každý nech si vyberie podlá svojho zmýšľania, či politickej príslušnosti.

Označkované slovo

.czechoslovakia.

(dĺžky 16) rozdelí na všetky podslová dĺžky 1, 2, ... Takto dostáva nasledujúce podslová:

dĺžky 1: `.`, `c`, `z`, `e`, ...
dĺžky 2: `.c`, `cz`, `ze`, ...
dĺžky 3: `.cz`, `zec`, `ech`, ...
...
dĺžky 15: `.czechoslovakia`, `czechoslovakia.`
dĺžky 16: `.czechoslovakia.`

Potom pre každé obdržané podslovo nájde medzi vzormi jeho „ohodnotený“ (celými čísla od 0 do 9) ekvivalent (nemusí vždy existovať). V našom prípade nájde v spomínanom súbore nasledujúce „ohodnotené“ podslová:

cz$_4$, $_2$ze, $_2$ch, $_3$cho$_2$, os$_4$l, $_4$lov, $_1$va

podľa ktorých priradí hodnotu každému znaku v rozdeľovanom slove spôsobom:

– znaku, za ktorým nenasleduje žiadne číslo – priradí 0,
– znaku, za ktorým sa stretne viac čísel – priradí najväčšie z nich.

Chceme napríklad určiť hodnotu znaku e v spomínanom slove. Podľa vzorky $_2$ch tam prislúcha 2 podľa $_3$cho$_2$ hodnota 3, ale podľa cze$_4$ hodnota 4. Výsledná hodnota znaku e je teda 4. Uvedeným spôsobom určíme hodnoty všetkých znakov, čím dostaneme (0 nepíšeme)

cz$_4$e$_4$cho$_2$s$_4$lo$_1$vakia.

A prekvapivý záver: TEX rozdelí slovo len za znakmi, ktoré majú nepárnu hodnotu. V našom prípade deliace miesto je teda jediné

czechoslo-vakia.

Pritom sa ešte kontrolujú momentálne nastavené hodnoty `\lefthyphenmin` (resp. `\righthyphenmin`) udávajúce kolko prvých (resp. posledných) znakov nesmie byť v slove rozdelených: pre anglíčtmu býva `\lefthyphenmin=2` a `\righthyphenmin=3`. (Toto sa ovšem nevzťahuje na slová rozdeľované pomocou `\hyphenation{...}`.)

Presvedčiť sa, že TEX skutočne rozdelí uvedené slovo daným spôsobom môžeme napísaním `\showhyphens{Czechoslovakia}` do svojho dokumentu a možné rozdelenia slova budú zapísané do súboru LOG. (Možno použiť i na celý paragraf.)

Viac informácií o niektorých jemnôstkach pri rozdeľovaní slov v TEXu môžu zvedavejší čitatelia nájsť v „biblii“ TEXbook [1, str. 449–455. Nájdete tam tiež poznámku o tom, že vytváranie „vzorkovníkov“ je dobre platená práca pre expertov...

Napriek tomu, že medzi spomínaných expertov nepatrím, každodenné „TEX-ovanie“ ma priviedlo k nutnosti vytvoriť aspoň nejakú verziu slovenských vzorov pre delenie. Na rozdiel od Lianga, ktorý základ svojich vzorov vyrobil z obrovského množstva slov a ich odvolení (k dispozícii mal Webster's Pocket Dictionary v elektronickej forme – zhruba 50 000 slov), použitá metóda je založená priamo na prepise gramatických pravidiel na rozdeľovanie slov. Možno niekomu poslúži ako inšpirácia na vytvorenie dokonalejšej verzie (prípadne poskytne mi nápady na vylepšenie), alebo ako ukážka toho, kadiaľ cesta nevedie.

Prvá fáza spočíva v prepise mechanických pravidiel rozdeľovania (t.j. pravidiel typu – ak je medzi dvoma samohláskami jedna spoluhláska, ...). Tieto dávajú nie vždy uspokojivé výsledky pri príponách a predponách, čo však je len chybou krásy a

nie gramatickou! V [2, str. 53] sa totiž píše, že ak si rozhranie medzi predponou, slovným základom a príponou neuvedomujeme, pripúšťa sa i mechanické rozdeľovanie slov. (Teda obe rozdelenia „náj-dem“ i „ná-jdem“ sú správne.) V prípade mechanického rozdeľovania slov vznikajú problémy so zloženými slovami, napr. rozdelenie Československo je hrubou (gramatickou) chybou.

Pre slovenčinu obe hodnoty `\lefthyphenmin` a `\righthyphenmin` nastavíme na 2. Podlá [2] totiž pri rozdeľovaní neoddeľujeme obyčajne koncovú a začiatočnú slabiku, ak obsahuje iba samohlásku.

Na začiatku zabránime deleniu dvojhlások `i2e`, `i2u`, `i2a` a spoluhlások `d2z`, `d2ž`, `c2h`. Potom pridáme každú samohlásku samostatne s 1:

`a1, 1, ä1, e1, é1, ...`

Tým si zabezpečíme správne delenie „pekných“ slov, v ktorých sa pravidelne strieda spoluhláska so samohláskou (slov typu „ko-lá-č“ máme ošetrené s `\righthyphenmin`). Dalšie pravidlo: ak sú medzi samohláskami dve spoluhlásky (r, l počítam vždy medzi spoluhlásky) prepíšeme do tvaru:

`zb1b, 2b1t, 2b1č. ...`

pre všetky možné dvojice spoluhlások, vrátane `dz, dž, ch`. *POZOR:* dvojice `2c1h, 2d1z, 2d1ž` musíme zo zrejmých dôvodov vyčiarknuť zo „vzorkovníka“.

Pridaním ďalšieho pravidla – výskyt troch spoluhlások medzi dvoma samohláskami pridáme do slovníka „vhodné“ trojice v ohodnotení:

..., `s3t2r` (ses-tra), `n3d2r (han-dra), n3s2k (pán-sky),` ...

Problém výberu takýchto trojíc je zložitejší ako v predchádzajúcom prípade dvojíc, lebo nemôžeme pridať všetky možné trojice (a to nielen kvôli kvantite). Napr. pridaním trojice `t3ĺ2k` by sme pokazili správne rozdelenie slova otlkať, ktoré dostaneme cez dvojice. Problém súvisí s neslabičnými `l, ĺ, r, ř`, ktoré v týchto prípadoch plnia funkciu samohlásky. Problém sa teda sústreduje na získanie zoznamu všetkých možných trojíc spoluhlások, vyskytujúcich sa v slovenských slovách. Nejaké štatistické výsledky možno nájsť v [3], rozhodne však nie sú uspokojivé.

Analogicky sa postupuje pri 4, 5, ... spoluhláskach, vyskytujúcich sa spolu. Problém vyhľadávania a ohodnocovania takýchto „zhlukov“ je podobného charakteru ako v predchádzajúcom prípade.

..., `r2s3t2v`, ...

V druhej fáze sa sústredujeme na predpony. Doterajší prepis zabraňuje rozdeľovaniu niektorých predpôn, napr. v slove „vystáť“ (po predpone vy-). Slovo bude rozdelené vys-táť. Nasleduje teda pridávanie predpôn

..., `.do3k4r, .do3k4l.,` ...

Zdôrazňujem, že uvedenie častí slovného základu (niekedy i viac než jednopísmenovej), je nutné. Ináč by sme pripustili nesprávne delenia, napr. delenie slova do-ktor.

Súčasťou tejto fázy je i pridávanie najfrekventovanejších častí zložených slov (v podstate čosi ako slovná predpona):

..., `5viac3h4` (viac-hlasný), `5česko5s4, .po4d3z, .štvor3r` ...

i s uvedením častí slovného z kladu z rovnakého dôvodu ako v predošlom.

V tretej fáze sa zameráme na prípony. Ich „pridanie“ do úvah spočíva v úprave ohodnotenia „zhlukov“, prípadne ich presnejšej špeciíikácii v spojení s väčšou časťou slovného základu.

Nakoniec nasleduje pridávanie pravidiel pre cudzie slová

`e1a2, u1 2, ...`

a výnimiek do `\hyphenation{médi-um, ...}`. Tejto častí však nebola zatiaľ venovaná veľká pozornosť.

Čo dodat na záver: Snáď zodpovedať aktuálnu otázku „Ako bude rozdelené (slovo) Československo?“ Vytvoreným slovenským slovníkom: Čes-ko-slo-ven-sko ...

Literatura

[1] Knuth D. E., The TEXbook, Addison-Wesley Pub. Comp. A AMS, 1986.

[2] Oravec J. a Laca V., Príručka slovenského pravopisu pre školy, SPN, Bratislava, 1976.

[3] Mistrík J., Frekvencia tvarov a konštrukcií v slovenčine, VEDA, Bratislava, 1985.

(*Jana Chlebíková*)

$\mathcal{AMS}$-LaTeX verze 1.0 a 1.1

V červenci roku 1990 nabídla American Mathematical Society veřejnosti první verzi $\mathcal{AMS}$-LaTeXu. Tato verze byla označena jako 1.0. V současné době je k dispozici verze 1.1, která se však od verze 1.0 liší zcela nepodstatně — ve verzi 1.1 byly pouze opraveny některé známé chyby. Dále budu proto jednoduše mluvit o $\mathcal{AMS}$-LaTeXu.

Pokud nepočítám příznivce PLAINu a odpůrce velkých balíků maker vůbec, stojí každý budoucí uživatel TEXu před volbou, zda používat $\mathcal{AMS}$-TEX nebo LaTeX (volba LaTeXu je samozřejmě ta nejlepší!). Pravděpodobné ve snaze tuto volbu ulehčit, byly činěny různé pokusy, jak $\mathcal{AMS}$-TEXpřiblížit LaTeXu nebo naopak. Výsledkem nejznámějších dvou z těchto pokusů jsou balíky LA$\mathcal{MS}$-TEX a $\mathcal{AMS}$-LaTeX. Můj soukromý dojem však je, že uživatelé LaTeXu nebo $\mathcal{AMS}$-TEXu se pouze upevní ve víře, že jejich volba byla ta správná. Je třeba zdůraznit, že LA$\mathcal{MS}$-TEX a $\mathcal{AMS}$-LaTeX není jedno a totéž a tyto systémy jsou dokonce nekompatibilní. Systém LA$\mathcal{MS}$-TEX je rozšíření $\mathcal{AMS}$-TEXu, což mimo jiné vychází ze skutečnosti, že autorem LA$\mathcal{MS}$-TEXu je M. D. Spivak (tj. autor $\mathcal{AMS}$-TEXu). LA$\mathcal{MS}$-TEX Přijímá mnohé rysy LaTeXu jako např. automatické číslování formulí, kapitol, sekcí atd., možnost křížových referencí, okolí pro snadnou tvorbu tabulek atd. Věc, kterou LA$\mathcal{MS}$-TEX od LaTeXu nepřebírá, je okolí `picture` (okolí pro tvorbu jednoduchých obrázků přímo ve zdrojovém textu).

Na druhé straně dává LA$\mathcal{MS}$-TEX velké možnosti při vytváření komutativních diagramů; kdo v textu potřebuje mnoho komutativních diagramů, tomu bych doporučil LA$\mathcal{MS}$-TEX. Další výhodou LA$\mathcal{MS}$-TEXu je poměrná uzavřenost systému a podrobná dokumentace (M. D. Spivak, který již dříve napsal The Joy of TEX — manuál $\mathcal{AMS}$-TEXu — napsal podrobný manuál pro LA$\mathcal{MS}$-TEX).

Na rozdíl od LA$\mathcal{MS}$-TEXu je $\mathcal{AMS}$-LaTeXrozšířením LaTeXu. Přitom rozšířením v tom smyslu, že základem je dokonce přímo soubor `latex.tex` bez nějakých úprav (předpokládá se, že bude použit soubor `latex.tex` verze 2.09 a to ne starší než z května 1986). Tato skutečnost mimo jiné znamená, že v $\mathcal{AMS}$-LaTeXu budou k dispozici všechny příkazy LaTeXu tak, jak je uživatel zná z Lamportova manuálu (pouze příkazy pro volbu fontů budou mít poněkud jiný význam jak se zmíním dále). Kdo